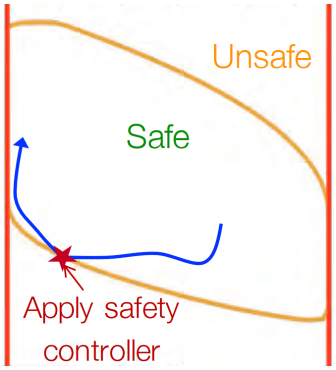
How to Achieve Optimality in Safe Reinforcement Learning?

Background

Many real-world applications such as autonomous driving, power system operation, or robot navigation require powerful decision-making tools. Reinforcement Learning (RL) has proven its potential to control complex systems from various applications by learning through interaction with the environment [4], [3]. However, most state-of-the-art RL algorithms have a significant disadvantage which prevents their deployment beyond simulated environments: They cannot guarantee fulfilling safety specifications, e.g., obstacle avoidance in autonomous driving. A simple yet effective approach for safeguarding RL algorithms is visualized below: If the action a_{RL}, which is proposed by the RL algorithm, steers the agent into an unsafe state, a safety controller is employed to modify a_{RL} so that the agent does not leave the safe set.



If the RL algorithm makes the agent leave the safe set, a safety controller is applied to keep the system inside the safe set (figure taken from: Melanie N. Zeilinger *Towards Safe Learing-Based Control*, around 2014).

The safety-preserving action a_{safe} can be obtained by projecting the (potentially unsafe) action a_{RL} onto a set of admissible actions $\mathcal{A}$ [1, Sec. 3.3.2]:

$$\begin{aligned} a_{\mathrm{safe}} = \arg\min_{a} \quad & ||a - a_{\mathrm{RL}}|| \\ \text{such that}: \quad & a \in \mathcal{A}. \end{aligned}$$

However, adjusting the action proposed by the RL algorithm may disrupt the learning process and result in suboptimal policies. Gros et al. [2] therefore derived corrections that ensure optimality in both Q-Learning and policy gradient methods despite safe action projections as described above. However, they did not provide practical examples showcasing the effects of their theoretical results.

Description

The goal of this thesis is to evaluate the influence of applying the corrections proposed in [2] to RL control of a dynamic system, e.g., an inverted pendulum. This includes extending an existing framework for safe RL of an inverted pendulum with the corrections of the learning process. Three different RL algorithms will be examined: Q-learning, deterministic policy gradient optimization and stochastic policy gradient optimization. For Q-learning, an additional goal is to analyze two different possibilities for integrating exploration into the learning process. One possibility is to randomly disturb the optimization problem formulation [5]. The other is to use epsilon-greedy action selection. The effect of the corrections should be evaluated empirically first. Optionally, it is then possible to derive theoretic bounds on the error introduced without the correction. Another extension would be to test the framework for more complex dynamic systems.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Hanna Krasowski, Hannah Markgraf, Lukas Schäfer

Type:
MA

Research area:
Safe Reinforcement Learning, Optimal Control

Programming language:
Python, Matlab

Required skills:
Basic understanding of cyber-physical systems, control theory, and reinforcement learning

Language:
English, German

For more information please contact us:

Phone:

E-Mail:
hannah.markgraf@tum.de

Website:
www.ce.cit.tum.de/en/air/home/

Tasks

- Familiarize with the corrections for safe Q-learning and policy gradient methods proposed in [2],
- Implement uncorrected safe Q-learning and policy gradient methods for an inverted pendulum,
- Implement proposed corrections for the different learning algorithms,
- Evaluate the performance difference between standard and corrected learning algorithms empirically,
- Implement and evaluate two different exploration schemes for Q-learning.

What We Offer

- Research in Machine Learning,
- Weekly meetings with your advisors,
- Flexible start and schedule for the thesis project,
- Thesis topics that will be tailored to your interest, and
- Good coffee in case you want to meet in Garching.

References

[1] Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe Learning in Robotics: From Learning-Based Control to Safe Reinforcement Learning. December 2021.

[2] Sebastien Gros, Mario Zanon, and Alberto Bemporad. Safe Reinforcement Learning via Projection on a Safe Set: How to Achieve Optimality? *IFAC-PapersOnLine*, 53(2):8076–8081, January 2020.

[3] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv:1509.02971 [cs, stat]*, July 2019. arXiv: 1509.02971.

[4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015.

[5] Mario Zanon and Sebastien Gros. Safe Reinforcement Learning Using Robust MPC. *IEEE Transactions on Automatic Control*, 66(8):3638–3652, August 2021. IEEE Transactions on Automatic Control.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems